Predictor Of Length Of Stay In ICUs

Team 4
Thomas Lever (tsl2b@virginia.edu)

November 26, 2024

Executive Summary

Dr. Jon Michel, Director of Data Science at UVA Health,

My model predicts Length Of Stay (LOS) and not whether a patient will die or not.
My model is based on vitals, demographics, and diseases.
Further directions include balancing data based on LOS, including medication information, and predicting whether a patient will die or not.
My feature matrix has 3,896 rows; columns with subject id, hospital admission id, ICU stay id, vitals, demographics, and diseases; and no null values.

We hope to offer a proof of concept for streamlining resource allocation in ICUs.

Context, Motivation, and Value

Intensive Care Units (ICU) play a pivotal role in modern healthcare by providing specialized care to critically ill patients. However, the unpredictable nature of patient admissions and varying Lengths Of Stay (LOS) pose significant challenges for resource allocation and operational efficiency. Accurately predicting the LOS in ICUs is crucial for optimizing use of beds, equipment, staff, and other resources. We are motivated to enhance healthcare by ensuring better planning, resource management, lower costs, and improved patient outcomes. By developing a reliable predictor of LOS in ICUs, we aim to provide valuable insights that can assist healthcare professionals in decision-making processes, ultimately leading to more efficient and effective critical care services.

Opportunity ICU Patient Outcome Predictor Will Address

Our predictor may assist doctors and nurses in allocating resources in an ICU.

What ICU Patient Outcome Predictor Will Do

Our ICU Patient Outcome Predictor will be a machine learning model trained on attributes of ICU patients that will predict LOS of ICU patients based on patient attributes.

Our model performs well.

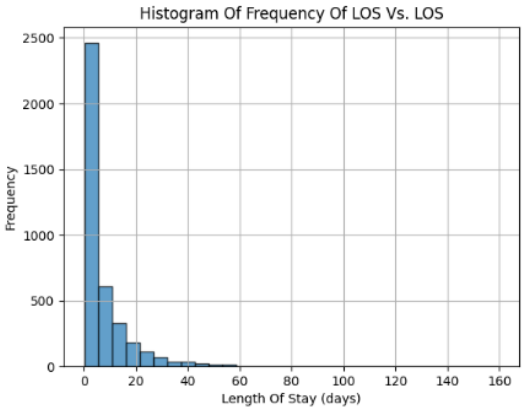
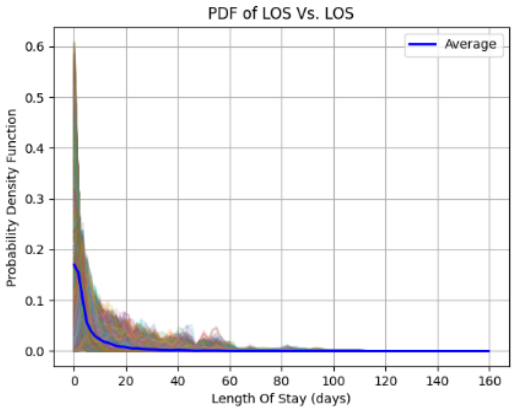
As described below, we trained a Random Survival Forest on observations of patient LOS and demographic information, vital signs, and indicators of whether those patients had conditions in the most 20 most common Disease Related Groups. The most important predictors were maximum respiratory rate, maximum heart rate, and minimum oxygen saturation pulseoxymetry.

Our model has an overall Mean Squared Error (MSE) of 49.622. MSE measures the average squared difference between actual and predicted expected LOS and quantifies the variance of prediction errors. Lower MSE indicates that predicted LOSs are closer to actual LOSs and the model performs better.

Our model has an overall Concordance Index of 0.821. Concordance Index measures the ability to correctly rank survival times and is the probability that, for a pair of randomly chosen observations, the observation with the shorter actual LOS is correctly predicted to be shorter. 0.5 indicates that our model is no better than a random guesser. 1.0 indicates that our model ranks perfectly. Our model does well.

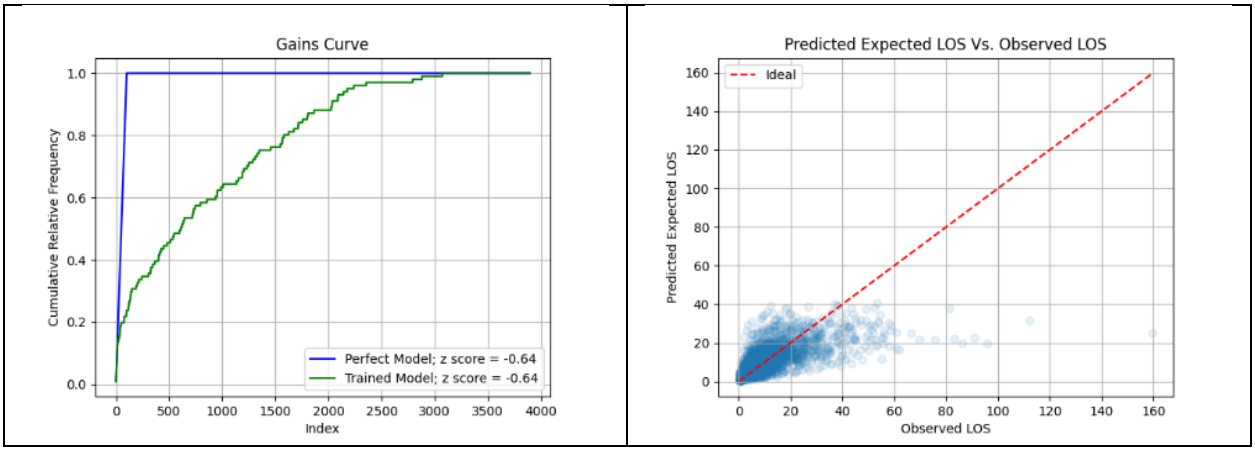
Below is a histogram of frequency of LOS vs. LOS for our entire data set. Also below is a graph of Probability Density Functions (PDF) of LOS vs. LOS for all patients, with a blue curve representing average PDF value. The two curves look very similar. Our model succeeds at predicting the probabilities of various LOSs for various patients. Following [3], the expected value of a PDF and the predicted expected survival time of a Survival Function (SF) from which a PDF is derived are both equal to

$$\mu = \int_{t=0}^{\infty} t\, PDF(t)\, dt = \int_{t=0}^{\infty} SF(t)\, dt$$

Histogram Of Frequency Of LOS Vs. LOS	PDF Of LOS Vs. LOS
	

Below are Gains Curves. Gains Curves may be interpreted as Receiver Operator Characteristic (ROC) Curves. Our Gains Curve indicates that our model has learned about the relationship between LOS and our predictors.

Gains Curves	Predicted Expected Vs. Observed LOS



We create a plot of predicted expected LOS vs. observed LOS. Fortunately, predicted expected LOS seems to be on the same order of magnitude as observed LOS. Unfortunately, predicted expected LOS seems to vary logarithmically with expected LOS. This might be a symptom of bias toward shorter LOSs as a result of there being few observations of long LOSs. This plot is arguably more linear for shorter LOSs. Further directions may be to balance our data set or apply a Band Aid of linearizing predictions.

We trained a Random Survival Forest.

Response

We developed a model to predict LOS of a patient in ICUs. LOS is given in table `icustays` in the MIMIC IV data set. Specifically, we predict SFs for patients and derive Cumulative Distribution Functions (CDF) and PDFs. A PDF is a distribution of likely LOSs. A CDF describes the probability that a random LOS is at most a specific duration. A SF describes the probability that a random LOS exceeds a specific duration. A SF is the complement of a CDF.

Predictors

Our model learns to predict LOS based on demographic information, vital signs, and Diagnosis Related Groups.

Model

Classification is often used to group patients into multiple classes, such as short, medium or long stay, or recovered or died. However, given that there are few examples of long stays or deaths in the hospital, the data is imbalanced, which biases any classification model. Predicting LOS using regression is more appropriate [2].

Survival analysis not only predicts numerical LOSs but also predicts curves relating to probability distributions of LOSs. Another advantage of survival analysis is that a survival model may be trained using censored data. A further direction is to work with curators of MIMIC IV

and/or the medical center to determine which observations in table `icustays` lack discharge time and/or LOS, and which observations had discharge times or LOSs censored to the end of an observation period.

We used a Random Survival Forest. We ignored all observations in table `icustays` with missing discharge time and/or LOS and assume that all patients were discharged from an ICU. We also ignored observations with missing IDs or predictor values.

We began a grid search for ideal values of the number of trees, maximum depth of trees, minimum number of samples required to split a node, minimum number of samples required to have a leaf node, and maximum number of features considered when splitting. Ranges for these values were [100, 1000], [[5, 50], None], [2, 10], [1, 10], and [`sqrt`, `log2`, [0.0, 1.0], **Z**, None]. Our featured model had values 100, None, 6, 3, and None.

We use 5 fold cross validation. For each of 5 subsets of data for testing, we train our model on the other 4 subsets of data and predict on the subset of data for testing. We compile all observed LOSs from all subsets of data for testing and their corresponding predicted expected LOSs.

We used the MIMIC IV version 3.0 data set.

Raw Data

We used the MIMIC (Medical Information Mart for Intensive Care) IV version 3.0 data set [1]. MIMIC is a large, publicly available database of deidentified electronic health records from the Beth Israel Deaconess Medical Center in Boston, MA. The data set contains information relating to patient measurements, diagnoses, procedures, and treatments.

MIMIC IV empowers researchers to advance understanding of care of patients and responses of patients to treatments. That being said, like MIMIC III and many other data sets, MIMIC IV contains imbalances in observations. For example, MIMIC IV's distribution of LOSs is skewed right. Age is relatively unimportant to our model. Race and gender are less important. It's possible that relatively few patients have certain ages, races, or genders and thus these demographic predictors are less important than they should be. It's possible that people of certain ages, races, or genders receive less quality care that results in LOSs that are too short or long, or even missing or inaccurate records. A further direction may be to balance observations of LOS and to ensure that all patients received comparable care.

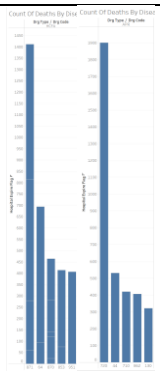
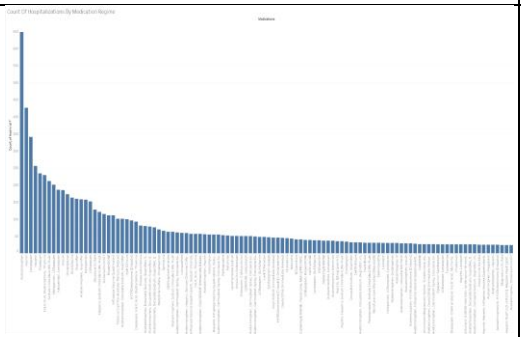
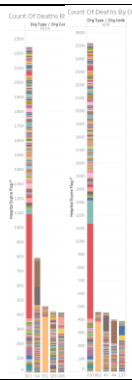
Data For Model

We considered training our predictor based on a table of patient information with 3895 rows and 62 columns. Each row corresponds to 1 patient, hospitalization, and stay in an ICU. There is 1 column of LOSs in an ICU. There are 3 columns of demographic information; namely, values of gender, age, and race. There are 7 groups each corresponding to a vital sign; namely, BMI, heart rate, diastolic blood pressure, systolic blood pressure, heart rate, oxygen saturation pulse oximetry, and temperature. Each group has 5 columns of minimum, first quartile, median, third

quartile, and maximum values for a given stay. There are 20 columns of indicators of whether or not a patient entered the hospital with a condition. For simplicity, our table of patient information and the tables from which it was derived have no missing values. A further direction might be to include columns in our table of indicators of whether patients received a treatment in a group of treatments.

Exploratory Data Analysis

"Count Of Deaths By Disease For Disease DRG Type HCFA" shows the top 5 counts of patient deaths by disease for disease DRG type HCFA. "Count Of Deaths By Disease For Disease DRG Type APR" shows the top 5 counts of patient deaths by disease for disease DRG type APR. A disease is identified in the data by one DRG type and one DRG code. For each disease type, a plurality of patients died from septicemia or severe sepsis… without more than 96 hours of Mechanical Ventilation (MV) and with Major Complication or Comorbidity (MCC).

Count Of Deaths By Disease For Disease DRG Types HCFA and APR	Count Of Hospitalizations By Medication Regime	Count Of Deaths By Disease And Treatment For Disease DRG Type HCFA and APR
		

"Count Of Hospitalizations By Medication Regime" shows counts of hospitalizations by medication regime. A medication regime is a set of alphabetized medications administered to a patient during a patient's hospitalization. For 650 hospitalizations only acetaminophen was administered.

"Count Of Deaths By Disease And Treatment For Disease DRG Type HCFA" represents the count of patient death by disease type and treatments for disease DRG type HCFA. "Count Of Deaths By Disease And Treatment For Disease DRG Type APR" represents the count of patient death by disease type and treatments for disease DRG type APR. A treatment is a solution to address one or more diseases. Solution in this case means "solution to a problem" as opposed to "chemical solution". For each disease type, a plurality of patients died from septicemia or severe sepsis… without more than 96 hours of Mechanical Ventilation (MV) and with Major Complication or Comorbidity (MCC). For this disease, a plurality of patients who died were treated with a treatment for two types of sepsis.

Our proof of concept is useful to ICUs.

Given that the Concordance Index of our model is greater than 0.8, we may be able to apply our model in a clinical setting. Our model may be useful in planning for use of resources and patient care for likely amounts of time. As mentioned below, we may wish to enhance our model by predicting LOS based on existing features and intended medications.

References

1. MIT Laboratory for Computational Physiology (2024). “Medical Information Mart for Intensive Care”. Accessed via https://mimic.mit.edu/ on 11/05/2024.
2. Wen et al. (2022). "Time-to-event modeling for hospital length of stay prediction for COVID-19 patients". Machine Learning with Applications. Volume 9. Jun 18, 2022. Accessed via https://pmc.ncbi.nlm.nih.gov/articles/PMC9213016/pdf/main.pdf on 10/17/2024.
3. Rodriguez, German (2024). “7. Survival Models”. Accessed via https://grodri.github.io/glms/notes/c7s1 on 11/05/2024.